Equilibrium Selection in Multi-Agent Policy Gradients via Opponent-Aware Basin Entry

author names withheld

Under Review for NExT-Game 2026

Abstract

Multi-agent policy-gradient methods have been shown to converge locally near stable Nash equilibria. Local convergence, however, does not determine which equilibrium is reached. We study this question through basin-entry probability with respect to a target set of equilibria selected by an external criterion, such as payoff dominance. For finite-unroll Meta-MAPG, we show that the update decomposes into ordinary policy gradient plus own-learning and peer-learning corrections, with controlled sampling noise and finite-unroll bias. We identify the peer-learning correction as the main equilibrium-selection mechanism: under a local alignment condition, the probability of entering the attraction region of the target stable-Nash set increases, relative to ordinary policy gradient. Because persistent correction may shift zero-update points of the original game, annealing the correction after entering the basin recovers ordinary policy-gradient dynamics and inherits local stable-Nash convergence guarantees. Experiments in Stag Hunt, iterated Prisoner's Dilemma, and preliminary neural-policy coordination environments support this basin-entry view, showing increased entry into cooperative basins under peer-aware updates.

1. Introduction

Giannou et al. [4] established that multi-agent policy gradient converges locally near stable Nash equilibria: from initialisations sufficiently close to one, the dynamics enter and stay. However, generic stochastic games admit multiple stable Nash equilibria with very different welfare, coordination, and safety properties, and local convergence is silent on which one learning reaches from an arbitrary initialisation. It is then natural to ask under what conditions the probability of entering the attraction region of a preferred equilibrium can be increased. We reframe this as a *finite-time basin-entry problem*: an external criterion W — payoff dominance, welfare, cooperation — designates a target set of stable Nash equilibria $\mathcal{N}_W^\star$, and we ask whether a learning algorithm can raise the probability that iterates enter their certified attraction regions within a finite horizon.

Steering toward a chosen basin requires reasoning about how a policy shapes the future learning of its peers. Meta-MAPG [5], originally developed for rapid adaptation, differentiates through both an agent's own future updates and those of every other agent. We reinterpret it as a basin-entry mechanism and show that its finite-unroll update decomposes into ordinary policy gradient plus own- and peer-learning corrections. The own-learning correction vanishes at Nash while the peer-learning correction need not, making the peer term the load-bearing equilibrium-selection mechanism. Under a local alignment condition, this correction expands the certified attraction region, increasing target-basin entry probability relative to plain policy gradient.

Because peer shaping displaces the fixed point from the Nash equilibria of the original game, we introduce a *shape-then-cool* schedule that anneals the correction after basin entry to recover ordinary policy-gradient dynamics and the local convergence guarantees of Giannou et al. [4]. In tabular Stag Hunt, Meta-MAPG expands the cooperative basin from 27.0% under PG to 42.6%, with ablations confirming the gain is carried by the peer term. The same dynamics holds in IPD where peer-aware methods reach 32–37% cooperative success against 8% for PG and own-only. Preliminary neural-policy experiments are noisier and tail-driven, but broadly support that peer-aware shaping alters equilibrium selection near basin boundaries.

2. Related Work

Policy-gradient convergence in stochastic games. Recent work establishes almost-sure local convergence of policy gradient to stable Nash equilibria [4], but these guarantees are conditional on attraction-region entry and do not characterise which basin is entered. We study the complementary finite-time question: whether opponent-aware updates can increase entry into a criterion-selected target basin.

Opponent-aware learning. Opponent-aware learning differentiates through anticipated opponent updates, beginning with LOLA [2] and followed by stable, consistent, and proximal refinements such as SOS, COLA, and POLA [6–8]. Meta-MAPG [5] extends this family to finite multi-step unrolls in multi-agent reinforcement learning. We use this line of work as the algorithmic basis for studying how own- and peer-learning corrections affect basin-entry geometry.

Mechanism-facing MARL experiments. Most evaluations of opponent-aware methods report final returns on neural benchmarks, leaving basin identity inferential. We therefore use classic tabular coordination games, Stag Hunt and iterated Prisoner's Dilemma, where basin geometry can be inspected directly, and treat neural benchmarks such as Overcooked, MPE, and PettingZoo IPD as preliminary transfer probes rather than the main evidence.

3. Meta-MAPG as Stochastic Approximation

We consider a finite discounted stochastic game with joint policy parameter ϕ. Ordinary policy gradient induces the joint vector field

$$v(\phi) := (\nabla_{\phi_i} V_i(\phi))_{i=1}^{N}.$$

Giannou et al. [4] show that ordinary PG converges locally once the iterate lies in a certified neighbourhood of a stable Nash equilibrium; full definitions are in Appendix A. Meta-MAPG [5] models peer learning explicitly by differentiating through agents' own and peers' future updates; we reinterpret this finite-unroll update as a basin-entry mechanism.

Following Kim et al. [5], we consider a finite-L unroll of inner-loop policy-gradient updates and differentiate agent i's objective with respect to its initial parameters. The meta-gradient of agent i with respect to its initial parameters decomposes as

$$\nabla_{\phi_0^i} V_{0:L}^i = \mathbb{E}\left[G_i(\tau_L) \left(\underbrace{\nabla_{\phi_0^i} \log \pi(\tau_0 \mid \phi_0^i)}_{\text{current}} + \sum_{\ell=0}^{L-1} \underbrace{\nabla_{\phi_0^i} \log \pi(\tau_{\ell+1} \mid \phi_{\ell+1}^i)}_{\text{own-learning}} + \sum_{\ell=0}^{L-1} \underbrace{\nabla_{\phi_0^i} \log \pi(\tau_{\ell+1} \mid \phi_{\ell+1}^{-i})}_{\text{peer-learning}} \right) \right],$$

where $G_i(\tau_L) = \sum_{t\geq 0} \gamma_{\text{disc}}^t R_i(s_t, a_t)$ is agent i's discounted return along τ_L.

We write the Meta-MAPG correction as

$$M_L(\phi) =: M_L^{\text{own}}(\phi) + M_L^{\text{peer}}(\phi), \tag{1}$$

where two terms collect the own- and peer-learning terms above. The full Meta-MAPG update is

$$F_L(\phi) := v(\phi) + M_L(\phi). \tag{2}$$

Finite unrolling introduces a truncation bias, which is included in the stochastic-approximation decomposition below. The two correction terms play different roles. The own-learning term captures how agent i's current policy shapes its own future adaptation; the peer-learning term captures how it influences the future updates of its peers. At a Nash equilibrium ϕ^*, each agent's policy gradient vanishes, so the own-learning correction vanishes at ϕ^*, though it may matter along the path. The peer-learning correction need not vanish at ϕ^*, so it can change the local vector field and shift the shaped fixed point. This makes it capable of reshaping basin-entry geometry, but also why cooldown is needed later.

Under bounded rewards, smooth policies, bounded likelihood-ratio moments, and summable finite-unroll bias, the sampled finite-unroll Meta-MAPG update admits the following stochastic-approximation decomposition.

Proposition 1 (Meta-MAPG as stochastic approximation) *The sampled finite-unroll Meta-MAPG update admits the decomposition*

$$g_n = v(\phi_n) + \lambda_n M_{L_n}(\phi_n) + \xi_{n+1} + b_n, \tag{3}$$

where $M_{L_n} = M_{L_n}^{\text{own}} + M_{L_n}^{\text{peer}}$, $\{\xi_{n+1}\}$ is a martingale-difference sequence with $\mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = 0$ and $\mathbb{E}[\|\xi_{n+1}\|^2 \mid \mathcal{F}_n] \leq \sigma^2$, and $\|b_n\| \leq C\beta_n$ for a deterministic sequence $\beta_n \to 0$.

The proof in Appendix B is a direct stochastic-approximation argument, obtaining the martingale-difference, variance, and finite-unroll bias bounds from the bounded-moment and truncation assumptions.

4. Opponent-Aware Basin Entry

Proposition 1 shows that finite-unroll Meta-MAPG can be viewed as ordinary PG plus a peer-aware perturbation, sampling noise, and finite-unroll bias. We now ask what this perturbation is useful for. Local convergence theory is conditional on being inside the right region to begin with, and does not characterise which basin learning enters from an arbitrary initialisation. We therefore study how the peer-learning correction can shape local attraction-region geometry and increase target-basin entry under an alignment condition.

For a criterion-selected target set $\mathcal{N}_W^\star$, let $\mathcal{B}_W := \bigcup_{\phi^* \in \mathcal{N}_W^\star} B_{r_{\text{att}}}(\phi^*)$ denote the union of certified local attraction regions; full definitions are in Appendix A.

4.1. Local Alignment Condition

The peer-learning correction improves basin entry when its local linearisation pulls iterates toward the target equilibrium ϕ^*. Define

$$\mu_M := -\lambda_{\max}\left(\frac{J_M + J_M^\top}{2}\right), \qquad J_M := DM_L(\phi^*), \tag{4}$$

The local alignment condition is $\mu_M > 0$; it means that, to first order, the peer correction contributes additional negative drift in the target direction.

4.2. Local Geometry Under Peer-Learning Correction

We state the result for a single target equilibrium $\phi^* \in \mathcal{N}_W^\star$; the same argument applies locally around each equilibrium in $\mathcal{N}_W^\star$. We characterise how adding the peer-learning correction λM_L to the base policy-gradient update changes the local geometry around ϕ^* in three concrete ways: it shifts the fixed point to a nearby point ϕ_λ^*, improves the rate at which iterates are pulled toward it, which we call drift improvement, and enlarges the certified attraction region when $\mu_M > 0$.

Proposition 2 (Basin geometry under peer learning) *Fix an SOS Nash equilibrium ϕ^* of v with drift constant $\mu > 0$, and suppose $v, M_L \in C^1$ near ϕ^*. Let $F_\lambda(\phi) := v(\phi) + \lambda M_L(\phi)$ for shaping strength $\lambda \geq 0$, and let L_λ denote the Lipschitz constant of DF_λ on $B_{\rho_\lambda}(\phi_\lambda^*)$.*

(i) ***Fixed-point shift.*** *F_λ has its zero at a shifted point ϕ_λ^* satisfying $\|\phi_\lambda^* - \phi^*\| = O(\lambda)$.*

(ii) ***Drift improvement.*** *When $\mu_M > 0$, F_λ contracts toward ϕ_λ^* with an improved rate:*

$$\langle F_\lambda(\phi),\, \phi - \phi_\lambda^* \rangle \leq -\tfrac{1}{2}(\mu + \lambda\mu_M)\|\phi - \phi_\lambda^*\|^2.$$

(iii) ***Certified basin expansion.*** *The SOS ball $B_{\rho_\lambda}(\phi_\lambda^*) := \{\phi \in \mathbb{R}^d : \|\phi - \phi_\lambda^*\| < \rho_\lambda\}$, with radius $\rho_\lambda := (\mu + \lambda\mu_M)/(2L_\lambda)$, has a strictly larger certified radius than the corresponding local PG certificate.*

The full proof is in Appendix C.

By enlarging the certified attraction region of ϕ^*, the peer-learning correction steers additional initialisations into $\mathcal{B}_W$. Provided the initialisation distribution charges the enlarged certified region with strictly greater probability than the original ball, this yields

$$p_{\text{entry}}(\text{Meta-MAPG}, T) > p_{\text{entry}}(\text{PG}, T). \tag{5}$$

Basin entry is therefore the right object to optimise for equilibrium selection, and the peer-learning correction is the mechanism that can improve it.

4.3. Cooldown

The same peer correction that improves basin entry also changes the vector field. Since $M_L^{\text{peer}}(\phi^*)$ need not vanish at a Nash equilibrium, the zeros of $v(\phi) + \lambda M_L(\phi)$ need not coincide with the zeros of $v(\phi)$. Constant opponent-aware shaping can therefore converge to a perturbed fixed point rather than to a Nash equilibrium of the original game. We use a shape-then-cool schedule: the peer correction is active during warm-up, and after a handoff time T it is either set to zero or annealed so that $\sum_{n \geq T} \alpha_n \lambda_n < \infty$.

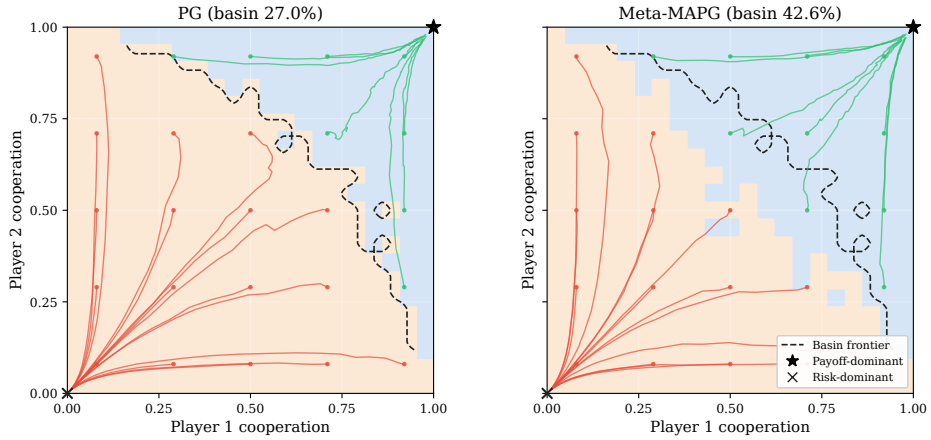

Figure 1: Stag Hunt basin geometry under PG (left) and full Meta-MAPG (right). The empirical cooperative basin expands from 27.0% to 42.6% on the 21×21 initialisation grid; the dashed contour is the PG separatrix reused in both panels as a fixed reference, not the Meta-MAPG separatrix.

Proposition 3 (Local convergence after cooldown) *Assume the stochastic-approximation decomposition in Proposition 1. If ϕ_T lies in the certified attraction region of an SOS Nash equilibrium ϕ^*, the post-handoff iterates remain in that local region, and the correction and finite-unroll bias are summable after handoff, then $\phi_n \to \phi^*$ almost surely.*

Thus peer-aware shaping is used only to improve target-basin entry; cooldown is used to recover convergence to a Nash equilibrium of the original game. The full proof is in Appendix D.

5. Experiments

The theory concerns certified local attraction regions, whereas the tabular experiments estimate empirical basin partitions by running learning from grids or paired initialisations. These objects need not coincide: the certified region is a local proof certificate, while the empirical basin is a global dynamical object. We use the maps as mechanism-facing evidence for the same basin-shaping effect. Stag Hunt makes the basin boundary drawable under Bernoulli policies; IPD provides a complementary tabular check of the peer-vs-non-peer split.

Figure 1 is the headline geometric result. On the same 21×21 initialisation grid, the cooperative basin in Stag Hunt grows from 27.0% under ordinary PG to 42.6% under full Meta-MAPG. The gained cells lie near the PG separatrix, consistent with the view that peer-aware shaping redirects trajectories across a basin boundary rather than merely accelerating learning.

Table 1 isolates the mechanism. Removing the peer term collapses performance to ordinary PG: own-only matches PG in both Stag Hunt and IPD. Retaining the peer term preserves the gain: peer-only matches full Meta-MAPG in Stag Hunt and reaches 32% cooperative success in IPD, against 37% for full Meta-MAPG. The empirical split is therefore peer-aware versus non-peer, not full Meta-MAPG versus peer-only, matching the decomposition in Section 3.

Two additional checks support this reading. First, the Stag Hunt alignment diagnostic points in the predicted direction: the first-update peer correction has mean cosine alignment 0.83 with the direction to (C, C) across the grid, and 0.84 on the 72 gained cells, with no lost cells. This

Method	Stag Hunt	IPD
PG	27% [19.3, 36.4]	8% [4.1, 15.0]
Own-only	27% [19.3, 36.4]	8% [4.1, 15.0]
Peer-only	42% [32.8, 51.8]	32% [23.7, 41.7]
Meta-MAPG	42% [32.8, 51.8]	37% [28.2, 46.8]

Table 1: Component ablation over 100 paired seeds, with Wilson 95% confidence intervals. Peer-aware methods separate from non-peer methods in both tabular games.

is a directional proxy, not a measurement of μ_M. Second, a Stag Hunt cooldown ablation retains the gain after annealing: PG reaches 31.25%, constant Meta-MAPG 41.25%, and shape-then-cool 42.50%. We therefore read cooldown as Nash-target insurance rather than as an empirically separated improvement at this horizon.

5.1. Neural-policy probe

We include a small neural-policy probe only to ask whether the peer-versus-non-peer pattern leaves any trace beyond tabular games. All arms share the same IPPO backbone, architecture, optimiser settings, and budget: `ippo`, `own_only`, `peer_only`, full `meta_mapg`, and fixed-time `handoff`. Each arm runs 25 paired seeds on Overcooked `forced_coordination`, three MPE tasks, and a PettingZoo IPD fallback; implementation details and full tables are in Appendices E–F.

The honest summary is that one environment carries the entire signal, and that signal is fragile. On Overcooked `forced_coordination`, `peer_only` has the highest mean final return (4.77), above `ippo` (3.10) and full `meta_mapg` (2.71), but every arm has median zero and the `peer_only` mean is driven by one seed reaching 117. Thus the neural result is consistent with peer-aware corrections occasionally surfacing a deeper coordination basin, but it is not evidence that Meta-MAPG outperforms IPPO. The remaining environments do not differentiate: MPE mostly saturates or overlaps across arms, and the IPD fallback collapses to defection.

6. Conclusion and Limitations

We have reinterpreted finite-unroll Meta-MAPG as a mechanism for equilibrium selection rather than rapid adaptation. Decomposing the meta-gradient exposes a structural asymmetry: the own-learning term vanishes at Nash while the peer-learning term need not, and only the latter can reshape the local geometry of the update field. Under a local alignment condition, this reshaping strictly enlarges the certified attraction region of a target stable Nash equilibrium, raising the probability of basin entry from an arbitrary initialisation; annealing the correction after entry recovers ordinary policy-gradient dynamics and inherits its local stable-Nash convergence guarantees. Basin entry, not asymptotic convergence, is the appropriate target for gradient-based equilibrium selection — and the peer-learning correction is the load-bearing mechanism through which it can be steered.

The guarantees are local and conditional: Proposition 2 requires the alignment condition $\mu_M > 0$, and Proposition 3 requires successful basin entry at the handoff time, together with standard regularity assumptions. The paper formalises criterion-selected basin entry but does not solve criterion choice — whether W should encode payoff dominance, welfare, fairness, or robustness is a separate question, and these desiderata can conflict. Many-agent scaling, the constructive role of own-learning in long-horizon or non-stationary settings, and scalable alignment diagnostics remain open.

References

[1] Arlington M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, Series A-I*, 28(1):89–93, 1964.

[2] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 122–130, 2018.

[3] Jakob Foerster, Gregory Farquhar, Maruan Al-Shedivat, Tim Rocktaschel, Eric Xing, and Shimon Whiteson. DiCE: The infinitely differentiable monte carlo estimator. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1524–1533, 2018.

[4] Angeliki Giannou, Kyriakos Lotidis, Panayotis Mertikopoulos, and Emmanouil V. Vlatakis-Gkaragkounis. On the convergence of policy gradient methods to nash equilibria in general stochastic games. In *Advances in Neural Information Processing Systems*, 2022.

[5] Dong-Ki Kim, Miao Liu, Matthew Riemer, Chuangchuang Sun, Marwa Abdulhai, Golnaz Habibi, Sebastian Lopez-Cot, Gerald Tesauro, and Jonathan P. How. A policy gradient algorithm for learning to learn in multiagent reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[6] Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktaschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. In *International Conference on Learning Representations*, 2019.

[7] Timon Willi, Alistair H. P. Letcher, Johannes Treutlein, and Jakob Foerster. COLA: Consistent learning with opponent-learning awareness. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23804–23831. PMLR, 2022.

[8] Stephen Zhao, Chris Lu, Roger B. Grosse, and Jakob Foerster. Proximal learning with opponent-learning awareness. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Appendix A. Problem Setup

A.1. Stochastic Games and Policies

We consider N-player finite discounted stochastic games. A game is specified by a tuple:

$$\mathcal{G} = (\mathcal{S},\, \mathcal{N},\, \{\mathcal{A}_i, R_i\}_{i\in\mathcal{N}},\, P,\, \gamma_{\text{disc}},\, \rho_0)\,,$$

where $\mathcal{S} = \{1, \dots, S\}$ is a finite state space, $\mathcal{N} = \{1, \dots, N\}$ is the set of agents, $\mathcal{A}_i = \{1, \dots, A_i\}$ is agent i's action set with joint space $\mathcal{A} = \prod_{i\in\mathcal{N}} \mathcal{A}_i$ and opponent space $\mathcal{A}_{-i} = \prod_{j\neq i} \mathcal{A}_j$, $R_i \colon \mathcal{S} \times \mathcal{A} \to [-1, 1]$ is agent i's reward function, $P(s' \mid s, \alpha)$ is the transition kernel, $\gamma_{\text{disc}} \in [0, 1)$ is the discount factor so that agent i's return is $\sum_{t\geq 0} \gamma_{\text{disc}}^t\, r_{i,t}$, and $\rho_0 \in \Delta(\mathcal{S})$ is the initial state distribution.

Each agent i plays a stationary Markovian policy $\pi_{\phi_i}(a_i \mid s)$, twice differentiable in its parameter $\phi_i \in \mathbb{R}^{d_i}$. The joint parameter $\phi = (\phi_i)_{i=1}^N \in \mathbb{R}^d$ is the single object all updates act on. Agent i's value function $V_{i,\rho}(\phi)$ is the expected discounted return under joint policy ϕ,

$$V_{i,\rho}(\phi) := \mathbb{E}_{p(\tau|\phi)}\left[\sum_{t\geq 0} \gamma_{\text{disc}}^t\, R_i(s_t, a_t)\right], \tag{6}$$

which each agent seeks to maximise. Because agents optimise their own $V_{i,\rho}$ simultaneously, no agent controls the full environment: every policy update by agent i changes the returns of all other agents. This interdependence is what makes equilibrium selection nontrivial.

The policy-gradient update for the joint parameter ϕ is

$$\phi_{n+1} = \phi_n + \alpha_n\, v(\phi_n), \qquad v(\phi) := \big(\nabla_{\phi_i} V_{i,\rho}(\phi)\big)_{i=1}^N, \tag{7}$$

where each component $\nabla_{\phi_i} V_{i,\rho}(\phi)$ is agent i's individual policy gradient and $\alpha_n > 0$ is the step size. We refer to this as ordinary policy gradient (PG).

A.2. Nash Equilibria and Certified Attraction Regions

The central solution concept in stochastic games is the Nash policy: a joint policy from which no agent can improve their return by unilaterally deviating. Formally,

Definition 4 (Nash Policy) *A joint policy $\phi^* = (\phi_i^*)_{i=1}^N$ is a Nash policy if for every agent $i \in \mathcal{N}$ and all $\phi_i' \in \mathbb{R}^{d_i}$,*

$$V_{i,\rho}(\phi_i^*, \phi_{-i}^*) \geq V_{i,\rho}(\phi_i', \phi_{-i}^*). \tag{8}$$

Fink [1] established that every finite discounted stochastic game admits at least one stationary Nash policy, so the solution concept is non-vacuous. However, existence alone does not guarantee that learning dynamics converge to a Nash policy. Some Nash policies are unstable under gradient updates, and learning algorithms may diverge away from them even when initialised nearby. We therefore work with stronger refinements that admit local stable-convergence guarantees.

Definition 5 (Stable Nash and SOS [4]) *A Nash policy ϕ^* is stable if*

$$\langle v(\phi),\, \phi - \phi^* \rangle < 0 \tag{9}$$

for all $\phi \neq \phi^*$ sufficiently close to ϕ^*, and second-order stationary (SOS) if there exist $\mu > 0$ and $\rho > 0$ such that

$$\langle v(\phi),\, \phi - \phi^* \rangle \leq -\mu \|\phi - \phi^*\|^2 \tag{10}$$

whenever $\|\phi - \phi^*\| \leq \rho$.

These conditions are nested — SOS implies stable, and stable implies Nash. A finite stochastic game can admit multiple stable Nash equilibria, each with its own certified attraction region. The certified attraction region $B_{r_{\mathrm{att}}}(\phi^*)$ is the set of initialisations from which Giannou et al. [4] guarantee convergence of ordinary PG.

Definition 6 (Certified attraction region) *Let ϕ^* be an SOS Nash equilibrium with constants $\mu > 0$ and $r_{\mathrm{att}} > 0$. The certified attraction region of ϕ^* is the open ball*

$$B_{r_{\mathrm{att}}}(\phi^*) := \{\phi \in \mathbb{R}^d : \|\phi - \phi^*\| < r_{\mathrm{att}}\} \tag{11}$$

on which the SOS drift condition holds.

Lemma 7 (Local PG convergence [4]) *If $\phi_0 \in B_{r_{\mathrm{att}}}(\phi^*)$ for an SOS Nash equilibrium $\phi^* \in \mathcal{E}_{\mathrm{stab}}$, and the step sizes satisfy $\sum_n \alpha_n = \infty$, $\sum_n \alpha_n^2 < \infty$, then $\phi_n \to \phi^*$ almost surely.*

Local convergence theory is therefore conditional on being inside the right region to begin with, and does not determine which equilibrium learning reaches from an arbitrary initialisation. A finite stochastic game can admit multiple stable Nash equilibria, and learning dynamics do not preferentially select among them. We formalise equilibrium selection by introducing an external criterion W — such as payoff dominance or social welfare — that identifies a target subset of stable Nash equilibria,

$$\mathcal{N}_W^\star := \{\phi^* \in \mathcal{E}_{\mathrm{stab}} : W(\phi^*) \geq \tau\}, \tag{12}$$

where $\mathcal{E}_{\mathrm{stab}}$ denotes the set of all stable Nash equilibria of the game. The target attraction region is the union of certified attraction regions of equilibria in $\mathcal{N}_W^\star$,

$$\mathcal{B}_W := \bigcup_{\phi^* \in \mathcal{N}_W^\star} B_{r_{\mathrm{att}}}(\phi^*). \tag{13}$$

The criterion W defines which runs count as successes; it does not modify any update step.

The central object of this paper is the probability that a learning algorithm enters the certified attraction region of the target set $\mathcal{N}_W^\star$ within a fixed horizon $T \geq 1$.

Definition 8 (Basin-entry probability) *For an algorithm A, horizon $T \geq 1$, and target set $\mathcal{N}_W^\star$, the basin-entry probability is*

$$p_{\mathrm{entry}}(A, T; W) := \Pr\big(\phi_T \in \mathcal{B}_W\big), \tag{14}$$

where ϕ_T is the iterate produced by A after T steps from an initial policy $\phi_0 \in \mathbb{R}^d$.

We note that the event $\phi_T \in \mathcal{B}_W$ is a finite-time event: it asks whether the iterate has reached a region where local convergence theory applies; what happens after entry is addressed in Section 4.3. This is distinct from asymptotic convergence, and it is also distinct from the event of entering any stable-Nash basin, which depends on game geometry and the initialisation distribution alone. Our central quantity, $p_{\text{entry}}(A, T; W)$, is the probability of entering the criterion-selected target basin $\mathcal{B}_W$, and whether opponent-aware updates improve target basin entry relative to plain PG,

$$p_{\text{entry}}(\text{Meta-MAPG}, T) > p_{\text{entry}}(\text{PG}, T), \tag{15}$$

and under what conditions on the peer-learning correction this holds.

Appendix B. Proof of Proposition 1: Meta-MAPG as Stochastic Approximation

We prove the statement for the joint stacked update. The proof is blockwise in the agents, and stacking the blocks gives the displayed vector decomposition.

At iteration n, let L_n denote the finite unroll length used by the Meta-MAPG estimator. If a fixed unroll length is used, this proof applies with $L_n \equiv L$. Let

$$\mathcal{F}_n := \sigma(\phi_0, \mathcal{Z}_0, \ldots, \mathcal{Z}_{n-1})$$

be the natural filtration generated by the initial parameter and all trajectory bundles used before iteration n, where $\mathcal{Z}_k$ denotes all sampling randomness used to construct the update at iteration k. Thus ϕ_n is $\mathcal{F}_n$-measurable, and the new trajectory bundle $\mathcal{Z}_n$ is sampled conditionally on $\mathcal{F}_n$ under the policies generated from ϕ_n.

For agent i, write the sampled update estimator in the form

$$g_{i,n} = \widehat{v}_{i,n} + \lambda_n \Big(\widehat{m}^{\text{own}}_{i,L_n,n} + \widehat{m}^{\text{peer}}_{i,L_n,n} \Big), \tag{16}$$

where $\widehat{v}_{i,n}$ is the sampled ordinary policy-gradient part and $\widehat{m}^{\text{own}}_{i,L_n,n}$, $\widehat{m}^{\text{peer}}_{i,L_n,n}$ are the sampled finite-unroll own-learning and peer-learning terms from the Meta-MAPG decomposition. Define the stacked vectors

$$g_n := (g_{i,n})_{i=1}^N, \qquad v(\phi_n) := \big(\nabla_{\phi_i} V_{i,\rho}(\phi_n)\big)_{i=1}^N,$$

and

$$M_{L_n}(\phi_n) := M^{\text{own}}_{L_n}(\phi_n) + M^{\text{peer}}_{L_n}(\phi_n),$$

where $M^{\text{own}}_{L_n}$ and $M^{\text{peer}}_{L_n}$ denote the deterministic finite-unroll corrections obtained by taking conditional expectations of the corresponding sampled own-learning and peer-learning estimators.

Step 1: Conditional mean of the sampled estimator.

Let

$$\bar{g}_n := \mathbb{E}[g_n \mid \mathcal{F}_n]$$

be the conditional mean of the sampled update. We first identify $\bar{g}_n$.

For the ordinary PG component, the likelihood-ratio identity gives, for each agent i,

$$\mathbb{E}[\widehat{v}_{i,n} \mid \mathcal{F}_n] = \nabla_{\phi_i} V_{i,\rho}(\phi_n). \tag{17}$$

The interchange of gradient and expectation is justified by (RC1)–(RC2): rewards are bounded, policies are twice differentiable, and the relevant score functions and Hessians are uniformly bounded on the local compact set K containing the iterates under consideration.

For the finite-unroll Meta-MAPG correction, applying the same likelihood-ratio identity recursively through the L_n inner-loop updates and using the chain rule gives

$$\mathbb{E}\Big[\widehat{m}^{\text{own}}_{i,L_n,n} + \widehat{m}^{\text{peer}}_{i,L_n,n} \,\Big|\, \mathcal{F}_n\Big] = M^{\text{own}}_{i,L_n}(\phi_n) + M^{\text{peer}}_{i,L_n}(\phi_n) + \delta_{i,n}, \tag{18}$$

where $\delta_{i,n}$ is the deterministic finite-unroll truncation residual. For an exact finite-unroll target, $\delta_{i,n} = 0$. When the finite unroll is used as an approximation to the untruncated meta-gradient, (RC4) gives a deterministic sequence $\beta_n \to 0$ and a constant $C < \infty$ such that, after stacking over agents,

$$\|\delta_n\| := \big\|(\delta_{i,n})_{i=1}^N\big\| \le C\beta_n. \tag{19}$$

The stronger form used later in Proposition 3 is $\sum_n \alpha_n\beta_n < \infty$.

Stacking (17) and (18), and absorbing the deterministic residual $\lambda_n\delta_n$ into b_n, we obtain

$$\bar{g}_n = v(\phi_n) + \lambda_n M_{L_n}(\phi_n) + b_n, \tag{20}$$

with

$$\|b_n\| \le C\beta_n. \tag{21}$$

If λ_n is kept explicit in the truncation residual, the constant C in (21) is enlarged using the assumed local boundedness of λ_n.

Step 2: Martingale-difference noise.

Define

$$\xi_{n+1} := g_n - \mathbb{E}[g_n \mid \mathcal{F}_n] = g_n - \bar{g}_n. \tag{22}$$

Then, by construction,

$$\mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[g_n - \bar{g}_n \mid \mathcal{F}_n] = 0. \tag{23}$$

Thus $\{\xi_{n+1}\}$ is a martingale-difference sequence with respect to $\{\mathcal{F}_{n+1}\}$.

The statement does not require the individual score terms to have zero mean. They generally do not: their conditional expectations are precisely the ordinary PG component and the Meta-MAPG correction terms. What is zero mean is the centered sampling error $g_n - \mathbb{E}[g_n \mid \mathcal{F}_n]$.

Step 3: Conditional second-moment bound.

By (RC3), used in its uniform-on-K form for the sampled finite-unroll likelihood-ratio estimator, there exists $\sigma^2 < \infty$ such that

$$\mathbb{E}\big[\|g_n - \mathbb{E}[g_n \mid \mathcal{F}_n]\|^2 \,\big|\, \mathcal{F}_n\big] \le \sigma^2 \qquad \text{for all } n. \tag{24}$$

Equivalently, using (22),

$$\mathbb{E}[\|\xi_{n+1}\|^2 \mid \mathcal{F}_n] \le \sigma^2. \tag{25}$$

This is the required conditional variance control. The uniformity in (24) is essential when L_n grows; it cannot in general be derived from the discount factor γ_{disc} alone, because γ_{disc} discounts environment time, not the length of the meta-learning unroll.

Step 4: Stochastic-approximation decomposition.

Combining (20) and (22), we obtain

$$g_n = \mathbb{E}[g_n \mid \mathcal{F}_n] + \xi_{n+1} = v(\phi_n) + \lambda_n M_{L_n}(\phi_n) + b_n + \xi_{n+1}. \tag{26}$$

Therefore the sampled finite-unroll Meta-MAPG update admits the stochastic approximation decomposition

$$g_n = v(\phi_n) + \lambda_n M_{L_n}(\phi_n) + \xi_{n+1} + b_n, \tag{27}$$

where

$$\mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = 0, \qquad \mathbb{E}[\|\xi_{n+1}\|^2 \mid \mathcal{F}_n] \le \sigma^2, \qquad \|b_n\| \le C\beta_n, \qquad \beta_n \to 0.$$

This proves Proposition 1. □

Appendix C. Proof of Proposition 2: Basin Geometry Under Peer Learning

We prove the three claims for the shaped vector field

$$F_\lambda(\phi) := v(\phi) + \lambda M_L(\phi),$$

where M_L is the finite-unroll Meta-MAPG correction. If the statement is specialised to the peer-learning correction only, the same proof applies with M_L replaced everywhere by M_L^{peer}.

The proof uses the local Lipschitz-Jacobian condition implicit in the definition of L_λ: for all sufficiently small λ, DF_λ is Lipschitz on the local neighbourhood under consideration. This holds, for example, if v and M_L are locally $C^{1,1}$ near ϕ^*.

Step 1: Consequences of the SOS condition

Let

$$A := Dv(\phi^*), \qquad B := DM_L(\phi^*),$$

and write

$$S_A := \frac{A + A^\top}{2}, \qquad S_B := \frac{B + B^\top}{2}.$$

We use $C \preceq D$ to denote Loewner order for symmetric matrices.

Since ϕ^* is an SOS Nash equilibrium of the ordinary policy-gradient field v, there exist $\mu > 0$ and $r > 0$ such that

$$\langle v(\phi), \phi - \phi^* \rangle \le -\mu \|\phi - \phi^*\|^2 \qquad \text{for all } \phi \in B_r(\phi^*). \tag{28}$$

First, (28) implies $v(\phi^*) = 0$. Indeed, for any unit vector u and sufficiently small $t > 0$,

$$\langle v(\phi^* + tu), tu \rangle \le -\mu t^2.$$

Dividing by t and sending $t \downarrow 0$ gives

$$\langle v(\phi^*), u \rangle \le 0.$$

Applying the same argument to $-u$ gives

$$\langle v(\phi^*), u \rangle \ge 0.$$

Hence $\langle v(\phi^*), u \rangle = 0$ for every unit vector u, and therefore $v(\phi^*) = 0$.

Next, substituting $\phi = \phi^* + tu$ into (28), using $v(\phi^*) = 0$, dividing by t^2, and sending $t \downarrow 0$ gives

$$u^\top Au \le -\mu\|u\|^2 \qquad \text{for every } u \in \mathbb{R}^d.$$

Since $u^\top Au = u^\top S_A u$, this is equivalent to

$$S_A \preceq -\mu I. \tag{29}$$

In particular, A is nonsingular: if $Au = 0$, then $u^\top Au = 0$, which contradicts $u^\top Au \le -\mu\|u\|^2$ unless $u = 0$.

The peer-alignment quantity is

$$\mu_M := -\lambda_{\max}\left(\frac{DM_L(\phi^*) + DM_L(\phi^*)^\top}{2}\right) = -\lambda_{\max}(S_B).$$

Thus $\mu_M > 0$ is equivalent to

$$S_B \preceq -\mu_M I. \tag{30}$$

Step 2: Fixed-point shift

Define

$$H(\phi, \lambda) := v(\phi) + \lambda M_L(\phi).$$

By Step 1,

$$H(\phi^*, 0) = v(\phi^*) = 0, \qquad D_\phi H(\phi^*, 0) = Dv(\phi^*) = A,$$

and A is nonsingular. Therefore, by the implicit function theorem, there exist $\bar\lambda > 0$, a neighbourhood U of ϕ^*, and a unique C^1 curve

$$[0, \bar\lambda] \ni \lambda \mapsto \phi^*_\lambda \in U$$

such that

$$F_\lambda(\phi^*_\lambda) = 0, \qquad \phi^*_0 = \phi^*.$$

Thus the zero of the shaped field persists locally under the perturbation λM_L.

Differentiating the identity

$$v(\phi^*_\lambda) + \lambda M_L(\phi^*_\lambda) = 0$$

at $\lambda = 0$ yields

$$Dv(\phi^*)\left.\frac{d\phi^*_\lambda}{d\lambda}\right|_{\lambda=0} + M_L(\phi^*) = 0.$$

Hence

$$\left.\frac{d\phi^*_\lambda}{d\lambda}\right|_{\lambda=0} = -A^{-1}M_L(\phi^*). \tag{31}$$

Since $\lambda \mapsto \phi^*_\lambda$ is C^1, its derivative is bounded after possibly reducing $\bar\lambda$. Therefore there exists a constant $C_{\text{shift}} > 0$ such that

$$\|\phi^*_\lambda - \phi^*\| \le C_{\text{shift}}\lambda \qquad \text{for all } 0 \le \lambda \le \bar\lambda.$$

Consequently,

$$\|\phi^*_\lambda - \phi^*\| = O(\lambda). \tag{32}$$

This proves part (i).

Step 3: Local drift improvement

Let

$$\theta_\lambda := \mu + \lambda\mu_M.$$

Combining (29) and (30), we obtain

$$\begin{aligned}\frac{DF_\lambda(\phi^*) + DF_\lambda(\phi^*)^\top}{2} &= \frac{Dv(\phi^*) + Dv(\phi^*)^\top}{2} + \lambda\frac{DM_L(\phi^*) + DM_L(\phi^*)^\top}{2} \\ &= S_A + \lambda S_B \\ &\preceq -(\mu + \lambda\mu_M)I = -\theta_\lambda I. \end{aligned} \tag{33}$$

Since $\phi^*_\lambda \to \phi^*$ as $\lambda \downarrow 0$, and DF_λ is continuous in ϕ locally, after reducing $\bar{\lambda}$ if necessary we also have

$$\frac{DF_\lambda(\phi^*_\lambda) + DF_\lambda(\phi^*_\lambda)^\top}{2} \preceq -\frac{3}{4}\theta_\lambda I \qquad \text{for all } 0 \le \lambda \le \bar{\lambda}. \tag{34}$$

Fix $0 < \lambda \le \bar{\lambda}$, and let

$$y := \phi - \phi^*_\lambda.$$

Since $F_\lambda(\phi^*_\lambda) = 0$, the fundamental theorem of calculus gives

$$F_\lambda(\phi^*_\lambda + y) = \int_0^1 DF_\lambda(\phi^*_\lambda + ty)y\, dt. \tag{35}$$

Taking the inner product with y, and using the fact that the skew-symmetric part of a matrix does not contribute to $y^\top Cy$, we obtain

$$\langle F_\lambda(\phi^*_\lambda + y), y\rangle = \int_0^1 y^\top \frac{DF_\lambda(\phi^*_\lambda + ty) + DF_\lambda(\phi^*_\lambda + ty)^\top}{2} y\, dt. \tag{36}$$

Let L_λ be a Lipschitz constant for DF_λ on $B_{\rho_\lambda}(\phi^*_\lambda)$. Then the symmetrised Jacobian is also L_λ-Lipschitz on this ball, so for every $t \in [0, 1]$,

$$\left\| \frac{DF_\lambda(\phi^*_\lambda + ty) + DF_\lambda(\phi^*_\lambda + ty)^\top}{2} - \frac{DF_\lambda(\phi^*_\lambda) + DF_\lambda(\phi^*_\lambda)^\top}{2} \right\|_{\text{op}} \le L_\lambda t\|y\|.$$

Combining this bound with (34) gives

$$y^\top \frac{DF_\lambda(\phi^*_\lambda + ty) + DF_\lambda(\phi^*_\lambda + ty)^\top}{2} y \le -\frac{3}{4}\theta_\lambda\|y\|^2 + L_\lambda t\|y\|^3.$$

Substituting into (36) yields

$$\begin{aligned}\langle F_\lambda(\phi^*_\lambda + y), y\rangle &\le \int_0^1 \left(-\frac{3}{4}\theta_\lambda\|y\|^2 + L_\lambda t\|y\|^3\right) dt \\ &= -\frac{3}{4}\theta_\lambda\|y\|^2 + \frac{L_\lambda}{2}\|y\|^3. \end{aligned} \tag{37}$$

Now define

$$\rho_\lambda := \frac{\theta_\lambda}{2L_\lambda} = \frac{\mu + \lambda\mu_M}{2L_\lambda}. \tag{38}$$

For every $\phi \in B_{\rho_\lambda}(\phi^*_\lambda)$, we have $\|y\| \le \rho_\lambda$, and therefore

$$\frac{L_\lambda}{2}\|y\|^3 \le \frac{\theta_\lambda}{4}\|y\|^2.$$

Using this in (37), we obtain

$$\langle F_\lambda(\phi^*_\lambda + y), y\rangle \le -\frac{1}{2}\theta_\lambda \|y\|^2.$$

Returning to $y = \phi - \phi^*_\lambda$, this becomes

$$\langle F_\lambda(\phi), \phi - \phi^*_\lambda\rangle \le -\frac{1}{2}(\mu + \lambda\mu_M)\|\phi - \phi^*_\lambda\|^2 \qquad \text{for all } \phi \in B_{\rho_\lambda}(\phi^*_\lambda). \tag{39}$$

This proves part (ii).

Step 4: Certified basin expansion

Equation (39) is precisely an SOS drift certificate for the shaped field F_λ around the shifted zero ϕ^*_λ, with drift constant

$$\frac{1}{2}(\mu + \lambda\mu_M)$$

on the ball

$$B_{\rho_\lambda}(\phi^*_\lambda), \qquad \rho_\lambda = \frac{\mu + \lambda\mu_M}{2L_\lambda}.$$

Thus $B_{\rho_\lambda}(\phi^*_\lambda)$ is a certified local attraction region for the shaped dynamics.

To make the strict radius comparison independent of the particular local Lipschitz upper bound chosen for each λ, compare ordinary PG and the shaped field using the same local certificate. Since v and M_L are locally $C^{1,1}$, after shrinking the neighbourhood of ϕ^* there exists a constant $L_{\mathrm{loc}} < \infty$ such that DF_λ is L_{loc}-Lipschitz on that neighbourhood for all sufficiently small λ. Using this common certificate, ordinary PG has certified radius

$$\rho_0 := \frac{\mu}{2L_{\mathrm{loc}}}, \tag{40}$$

while the shaped field has certified radius

$$\widetilde{\rho}_\lambda := \frac{\mu + \lambda\mu_M}{2L_{\mathrm{loc}}}. \tag{41}$$

If $\mu_M > 0$ and $\lambda > 0$, then

$$\widetilde{\rho}_\lambda - \rho_0 = \frac{\lambda\mu_M}{2L_{\mathrm{loc}}} > 0. \tag{42}$$

Therefore the local SOS certificate for the shaped field has strictly larger radius than the corresponding local SOS certificate for ordinary PG.

If the notation r_{att} in the main text denotes this ordinary local certificate, namely $r_{\mathrm{att}} = \rho_0$, then

$$\widetilde{\rho}_\lambda > r_{\mathrm{att}} \qquad \text{for every sufficiently small } \lambda > 0.$$

More generally, if r_{att} is a pre-existing certified radius, the same strict comparison holds whenever

$$r_{\mathrm{att}} \le \rho_0.$$

Finally, since $\widetilde{\rho}_\lambda > \rho_0$, the shaped certified ball $B_{\widetilde{\rho}_\lambda}(\phi^*_\lambda)$ cannot be contained in the ordinary certified ball $B_{\rho_0}(\phi^*)$. Hence the shaped dynamics certify additional local initialisations around the target equilibrium. This proves part (iii), and completes the proof of Proposition 2. □

Appendix D. Cooldown and Local Nash Convergence

Proposition 2 certifies that the peer correction enlarges the target basin during warm-up. What it does not address is what happens after entry: the asymptotic target. This gap matters, and the reason is the same asymmetry that made the peer-learning term useful in the first place. Because the peer correction does not vanish at Nash, holding it fixed permanently displaces the zero-update point — the stationary points of

$$v(\phi) + \lambda M_L(\phi) = 0 \tag{43}$$

do not coincide with those of v, and Proposition 2(i) quantifies the displacement as $\|\phi^*_\lambda - \phi^*\| = O(\lambda)$. The correction that steered the iterate into the right basin, if left on, leads Meta-MAPG to a perturbed equilibrium rather than a Nash equilibrium of the original game.

A shape-then-cool schedule resolves this. During warm-up,

$$\phi_{n+1} = \phi_n + \alpha_n\big(\hat{v}_n + \lambda_n \widehat{M}_{L_n,n}\big), \tag{44}$$

with λ_n non-negligible; after a handoff time T, λ_n is either set to zero or annealed so that $\sum_{n\geq T} \alpha_n\lambda_n < \infty$. The handoff can be fixed in advance or driven by an observable statistic such as cooperation rate or task return, and does not require an oracle that identifies basin entry during training. Proposition 3 below formalises this: provided the iterate has entered $B_{\mathrm{r_{att}}}(\phi^*)$ by time T, annealing the correction recovers ordinary policy-gradient dynamics and inherits its local stable-Nash convergence guarantee.

We now prove Proposition 3.

Step 1: Post-handoff dynamics and local bounds

Let

$$B := B_{r_{\mathrm{att}}}(\phi^*)$$

be the certified attraction region of the SOS Nash equilibrium ϕ^*. On this region, the SOS drift condition gives a constant $\mu > 0$ such that

$$\langle v(\phi),\, \phi - \phi^* \rangle \leq -\mu\|\phi - \phi^*\|^2, \qquad \phi \in B. \tag{45}$$

Work on the successful-entry event

$$\mathcal{E}_T := \{\phi_T \in B\}.$$

By the assumption of Proposition 3, the post-handoff iterates remain in the local certified region on which (45) holds.

For $n \geq T$, Proposition 1 gives the sampled update

$$\phi_{n+1} = \phi_n + \alpha_n\Big(v(\phi_n) + \lambda_n M_{L_n}(\phi_n) + \xi_{n+1} + b_n\Big), \tag{46}$$

where

$$M_{L_n} = M^{\mathrm{own}}_{L_n} + M^{\mathrm{peer}}_{L_n}, \qquad \mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = 0, \qquad \mathbb{E}[\|\xi_{n+1}\|^2 \mid \mathcal{F}_n] \leq \sigma^2,$$

and

$$\|b_n\| \leq C\beta_n.$$

The finite-unroll bias is summable after multiplication by the outer step size:

$$\sum_{n\geq T} \alpha_n \beta_n < \infty. \tag{47}$$

The proposition allows either hard cooldown, $\lambda_n = 0$ for all $n \geq T$, or annealed cooldown with

$$0 \leq \lambda_n \leq \bar{\lambda} < \infty, \qquad \sum_{n\geq T} \alpha_n \lambda_n < \infty.$$

The hard-cooldown case is included in the same notation by taking $\bar{\lambda} = 0$.

By the local boundedness assumptions in (RC2) and the proposition's local-region assumption, there exist finite constants $B_v, B_M < \infty$ such that, for all post-handoff iterates,

$$\|v(\phi_n)\| \leq B_v, \qquad \|M_{L_n}(\phi_n)\| \leq B_M, \qquad n \geq T. \tag{48}$$

Define

$$e_n := \phi_n - \phi^*, \qquad X_n := \|e_n\|^2,$$

and collect the deterministic post-handoff perturbations into

$$q_n := \lambda_n M_{L_n}(\phi_n) + b_n. \tag{49}$$

Then (46) can be written as

$$e_{n+1} = e_n + \alpha_n\big(v(\phi_n) + q_n + \xi_{n+1}\big). \tag{50}$$

Step 2: One-step Lyapunov inequality

Expanding $X_{n+1} = \|e_{n+1}\|^2$ using (50) gives

$$\begin{aligned} X_{n+1} = X_n &+ 2\alpha_n\langle e_n, v(\phi_n) + q_n + \xi_{n+1}\rangle \\ &+ \alpha_n^2\|v(\phi_n) + q_n + \xi_{n+1}\|^2. \end{aligned} \tag{51}$$

Taking conditional expectation with respect to $\mathcal{F}_n$ and using $\mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = 0$, we obtain

$$\begin{aligned} \mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n &+ 2\alpha_n\langle e_n, v(\phi_n)\rangle + 2\alpha_n\langle e_n, q_n\rangle \\ &+ \alpha_n^2\mathbb{E}\big[\|v(\phi_n) + q_n + \xi_{n+1}\|^2 \mid \mathcal{F}_n\big]. \end{aligned} \tag{52}$$

The SOS drift condition (45) gives

$$2\alpha_n\langle e_n, v(\phi_n)\rangle \leq -2\mu\alpha_n X_n. \tag{53}$$

For the perturbation term, Young's inequality gives

$$2\alpha_n\langle e_n, q_n\rangle \leq \mu\alpha_n X_n + \frac{\alpha_n}{\mu}\|q_n\|^2. \tag{54}$$

Combining (53) and (54),

$$2\alpha_n\langle e_n, v(\phi_n)\rangle + 2\alpha_n\langle e_n, q_n\rangle \leq -\mu\alpha_n X_n + \frac{\alpha_n}{\mu}\|q_n\|^2. \tag{55}$$

It remains to bound the second-moment term in (52). By (48), the boundedness of λ_n, the bound $\|b_n\| \le C\beta_n$, and $\sup_{n\ge T}\beta_n < \infty$, there exists $B_q < \infty$ such that $\|q_n\| \le B_q$ for all $n \ge T$. Therefore, using $\mathbb{E}[\|\xi_{n+1}\|^2 \mid \mathcal{F}_n] \le \sigma^2$,

$$\begin{aligned}\mathbb{E}\big[\|v(\phi_n) + q_n + \xi_{n+1}\|^2 \mid \mathcal{F}_n\big] &\le 3\|v(\phi_n)\|^2 + 3\|q_n\|^2 + 3\mathbb{E}[\|\xi_{n+1}\|^2 \mid \mathcal{F}_n] \\ &\le 3B_v^2 + 3B_q^2 + 3\sigma^2 =: C_1. \end{aligned} \tag{56}$$

Substituting (55) and (56) into (52), we get

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \le (1 - \mu\alpha_n)X_n + \frac{\alpha_n}{\mu}\|q_n\|^2 + C_1\alpha_n^2. \tag{57}$$

Step 3: Summability of the perturbation terms

From the definition of q_n in (49),

$$\|q_n\| \le \lambda_n\|M_{L_n}(\phi_n)\| + \|b_n\| \le B_M\lambda_n + C\beta_n.$$

Thus

$$\begin{aligned}\sum_{n\ge T}\alpha_n\|q_n\|^2 &\le 2B_M^2\sum_{n\ge T}\alpha_n\lambda_n^2 + 2C^2\sum_{n\ge T}\alpha_n\beta_n^2 \\ &\le 2B_M^2\bar{\lambda}\sum_{n\ge T}\alpha_n\lambda_n + 2C^2\bar{\beta}\sum_{n\ge T}\alpha_n\beta_n \\ &< \infty, \end{aligned} \tag{58}$$

where

$$\bar{\beta} := \sup_{n\ge T}\beta_n < \infty.$$

The first term is finite by the cooldown condition $\sum_{n\ge T}\alpha_n\lambda_n < \infty$, and it is identically zero in the hard-cooldown case $\lambda_n = 0$ for all $n \ge T$. The second term is finite by (47). Also,

$$\sum_{n\ge T}\alpha_n^2 < \infty$$

by the step-size assumption.

Define

$$r_n := \frac{\alpha_n}{\mu}\|q_n\|^2 + C_1\alpha_n^2.$$

Then (58) and $\sum_{n\ge T}\alpha_n^2 < \infty$ imply

$$\sum_{n\ge T} r_n < \infty. \tag{59}$$

Consequently, the Lyapunov recursion (57) takes the form

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \le (1 - \mu\alpha_n)X_n + r_n, \qquad \sum_{n\ge T} r_n < \infty. \tag{60}$$

Step 4: Robbins–Siegmund convergence

Since $\sum_{n\geq T}\alpha_n^2 < \infty$, we have $\alpha_n \to 0$. Hence, after discarding finitely many post-handoff iterates if necessary, we may assume $0 \leq \mu\alpha_n \leq 1$. This finite prefix does not affect the limit.

Applying the Robbins–Siegmund almost-supermartingale lemma to (60), we obtain that X_n converges almost surely to a finite random variable X_∞ and that

$$\sum_{n\geq T}\alpha_n X_n < \infty \qquad \text{almost surely.} \tag{61}$$

Because

$$\sum_{n\geq T}\alpha_n = \infty,$$

the only possible limit is $X_\infty = 0$. Indeed, if $X_\infty > 0$ on an event of positive probability, then on that event there exists N such that $X_n \geq X_\infty/2$ for all $n \geq N$, which would imply

$$\sum_{n\geq T}\alpha_n X_n \geq \sum_{n\geq N}\alpha_n X_n \geq \frac{X_\infty}{2}\sum_{n\geq N}\alpha_n = \infty,$$

contradicting (61). Therefore

$$X_n = \|\phi_n - \phi^*\|^2 \longrightarrow 0 \qquad \text{almost surely.}$$

Equivalently,

$$\phi_n \to \phi^* \qquad \text{almost surely.}$$

Thus, after successful entry into $B_{r_{\text{att}}}(\phi^*)$, either hard cooldown or summable annealing removes the asymptotic effect of the Meta-MAPG correction. The limiting dynamics recover ordinary PG and converge to the original SOS Nash equilibrium ϕ^*. □

Appendix E. Neural-Policy Implementation Details

E.1. IPPO backbone

All five arms share an Independent PPO (IPPO) backbone. Each agent i maintains a separate actor π_{θ_i} and critic V_{ϕ_i}. Rollouts of length $L = 256$ (or 400 for Overcooked) are collected per training iteration. Advantages are computed via Generalised Advantage Estimation with $\gamma_{\text{disc}} = 0.99$ and $\lambda_{\text{GAE}} = 0.95$. The PPO clipped surrogate is used with $\epsilon_{\text{clip}} = 0.2$, entropy coefficient 0.01, and gradient-norm clip 0.5. Each rollout uses 4 PPO epochs with minibatch size 256. Actor learning rate is $3 \cdot 10^{-4}$; critic learning rate is $1 \cdot 10^{-3}$.

E.2. Correction terms

The own- and peer-learning corrections are estimated using the DiCE estimator [3], which provides a differentiable Monte Carlo surrogate for higher-order policy gradients through a learning step. With inner step $\eta_{\text{inner}} = 0.1$ and advantage estimate $\hat{A}_i$, the per-agent correction terms are

$$\Delta_i^{\text{own}} = \nabla_{\theta_i}\log\pi_{\theta_i}(a_i \mid s)\cdot\hat{A}_i, \qquad \Delta_i^{\text{peer}} = \nabla_{\theta_i}\log\pi_{\theta_{-i}}(a_{-i} \mid s)\cdot\hat{A}_i.$$

Each arm activates a different subset: `ippo` uses Δ_{PG} only; `own_only` adds Δ^{own}; `peer_only` adds Δ^{peer}; `meta_mapg` adds both; `handoff` runs as `meta_mapg` until step T_{warm} and then disables corrections (revert to `ippo`). All corrections are norm-clipped to prevent blow-up:

$$\|\Delta_{\text{corr}}\|_2 \leq c \cdot \|\Delta_{\text{PG}}\|_2, \qquad c = 1.0.$$

This is a PPO-compatible practical proxy for the decomposition in Section 3, not a theorem-level object: the truncated correction is a deep-MARL surrogate that preserves the own-vs-peer distinction while remaining numerically tractable at neural scale.

E.3. Paired-seed protocol

For each seed value $\in \{0, 1, \ldots, 24\}$, all five arms initialise actor and critic weights from the same PRNG state. Differences in final return at fixed seed are therefore attributable to the algorithmic correction, not initialisation noise. The same protocol is used for all five environments.

Benchmark	Env id	Arms	Seeds	Steps	T_{warm}	τ
MPE	`simple_spread`	5	25	750 k	200 k	-25.0
MPE	`simple_reference`	5	25	750 k	200 k	-22.0
MPE	`simple_speaker_listener`	5	25	750 k	200 k	-10.0
Overcooked-AI	`forced_coordination`	5	25	1 M	500 k	3.0
Melting Pot[†]	`prisoners_dilemma_in_the_matrix`	4	25	1 M	400 k	0.5

Table 2: Per-environment configuration. T_{warm} applies only to the `handoff` arm; τ is the basin-entry threshold used for learner counts. [†] The `own_only` arm is omitted on this benchmark; the `dm_meltingpot` substrate was unavailable in the runtime environment, and we use the PettingZoo Iterated Prisoner's Dilemma fallback as a matrix-style proxy.

E.4. Sparse-reward fix in Overcooked

The Overcooked-AI environment wrapper originally returned only the sparse soup-delivery reward ($+20$), discarding the per-step shaped reward (onion pickup, dish handling). Without shaping, vanilla PPO cannot learn `forced_coordination` within 1 M steps. We patched the wrapper to add $0.5 \cdot \text{shaped}_i$ to the per-agent reward (the Overcooked-AI default coefficient), with the sparse soup reward split equally between the two cooks. The patch was applied uniformly across all reported Overcooked results.

E.5. Pilot-threshold cascade

The pilot stage ran post-patch and produced thresholds $\tau = 0.0$ for Overcooked and the IPD fallback. The full sweep used these stale thresholds, so success-fraction summaries on those two environments report effectively any-non-zero return as success. The honest signal is in mean final return and per-seed distribution, which are what we report in the main text and in Appendix F.

E.6. Evaluation and compute

Every $\sim$ 25 k environment steps we run 50 evaluation episodes with greedy (ARGMAX) action selection and report mean episode return. A seed counts as a learner if its final-checkpoint mean return crosses the benchmark-specific threshold τ. Greedy evaluation underestimates stochastic-policy performance; this affects all arms equally and is unlikely to change relative orderings. All training was performed on $4\times$ NVIDIA Tesla V100-SXM2-32GB GPUs with 12 parallel workers via `multiprocessing.spawn`; total compute is approximately 350 GPU-hours.

Appendix F. Full Neural-Sweep Tables

This appendix provides the complete per-environment final-return distributions that bound the scope of the main-text Overcooked claim. We organise environments from least- to most-differentiating.

F.1. MPE `simple_spread`: saturated

Arm	n	mean	std	min	p25	median	p75	max
`ippo`	25	-22.02	0.75	-23.87	-22.55	-22.08	-21.43	-20.48
`own_only`	25	-21.85	0.55	-22.80	-22.30	-21.86	-21.35	-20.65
`peer_only`	25	-21.81	0.77	-23.46	-22.18	-21.68	-21.48	-20.01
`meta_mapg`	25	-21.80	0.55	-23.13	-22.21	-21.74	-21.47	-20.67
`handoff`	25	-21.86	0.66	-23.05	-22.12	-21.83	-21.69	-19.75

Table 3: `simple_spread`, $\tau = -25$. All arms reach $25/25$ learners; mean returns differ by less than within-arm standard deviation. Median first-hit step is identical across arms.

The 3-agent / 3-landmark cooperative navigation task is fully saturated. This benchmark serves as a control: opponent-aware corrections do not hurt on easy cooperative tasks, but neither do they help.

F.2. MPE `simple_reference`: weakly suggestive

Arm	n	mean	std	min	p25	median	p75	max
`ippo`	25	-21.98	3.87	-36.20	-22.14	-20.81	-19.74	-18.79
`own_only`	25	-22.68	4.96	-38.19	-22.60	-21.02	-19.90	-19.24
`peer_only`	25	-22.21	3.36	-31.39	-24.07	-21.06	-19.64	-18.85
`meta_mapg`	25	-22.21	4.21	-37.70	-22.08	-21.16	-19.78	-19.02
`handoff`	25	-22.87	4.94	-42.38	-22.93	-21.26	-20.29	-18.95

Table 4: `simple_reference`, $\tau = -22$. Larger seed variance reveals tail behaviour; learner counts span 16–$18/25$ with overlapping Wilson intervals.

The 2-agent communication task introduces real seed variance (std $\approx$ 4 vs 0.5–0.8 on `simple_spread`). `peer_only` exhibits the smallest within-arm dispersion (std $=$ 3.36);

handoff shows the worst-case tail seed (min $= -42.38$). Wilson intervals on learner counts overlap heavily and the nominal ordering does not match the mean-return ordering.

F.3. MPE simple_speaker_listener: equal failure

Arm	n	mean	std	min	p25	median	p75	max
ippo	25	-25.81	1.96	-29.87	-27.44	-25.46	-24.22	-22.10
own_only	25	-25.82	2.05	-30.73	-27.71	-25.21	-24.37	-21.82
peer_only	25	-25.77	1.95	-29.84	-27.15	-25.65	-24.36	-21.58
meta_mapg	25	-25.88	2.04	-29.72	-27.45	-25.83	-24.26	-21.71
handoff	25	-25.64	2.01	-28.97	-27.71	-25.54	-24.13	-21.81

Table 5: simple_speaker_listener, $\tau = -10$. All arms fail to cross threshold within 750 k env-steps; final returns are near-identical across arms.

The asymmetric speaker-listener communication task is uniformly failed across all five arms. All arms produce near-identical statistics; differences are about one seed wide.

F.4. Overcooked forced_coordination: super-bimodal

Arm	n	mean	std	min	p25	median	p75	max
ippo	25	3.10	9.23	0.00	0.00	0.00	0.00	34.50
own_only	25	0.06	0.20	0.00	0.00	0.00	0.00	0.75
peer_only	25	4.77	22.91	0.00	0.00	0.00	0.00	117.00
meta_mapg	25	2.71	8.76	0.00	0.00	0.00	0.00	33.00
handoff	25	0.21	0.58	0.00	0.00	0.00	0.00	2.25

Table 6: forced_coordination, $\tau = 3.0$. Median final return is 0.00 for every arm; distributions are super-bimodal (most seeds fail entirely, rare learners reach high returns). The mean is dragged by tail learners.

forced_coordination is the strongest coordination test in the suite: the two cooks are physically separated by an interior counter, soup cannot be made by either alone, and partial credit is awarded only via shaped reward (onion pickup, dish handling, soup delivery). Coordinated soup delivery requires the agents to converge on a sequential convention. The full per-arm distribution is asymmetric across arms in two respects. peer_only produced one seed reaching 117 — a fully-synchronised partnership delivering multiple soups per evaluation episode — with all 24 remaining seeds at zero or near-zero. ippo produced four learners between 3 and 34.5, no deeper than that. own_only produced no learners and a maximum of 0.75 across all 25 seeds; handoff produced no learners and a maximum of 2.25. These distributions motivate the main-text reading of "fragile, tail-driven differentiating signal".

Arm	n	mean	std	min	p25	median	p75	max
ippo	25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
peer_only	25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
meta_mapg	25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
handoff	25	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 7: prisoners_dilemma_in_the_matrix__repeated (PettingZoo IPD fallback), $\tau = 0.5$. All arms collapse to mutual defection; mean and std are zero across arms.

F.5. PettingZoo IPD fallback: equilibrium collapse

The PettingZoo IPD fallback equilibrium-collapses to mutual defection, the Nash equilibrium for short-horizon IPD without communication or memory. Median first-hit on this benchmark is the very first eval checkpoint (step 512); seeds transiently exhibit cooperative spikes in the early random-policy regime, then descend to defection. This benchmark is uninformative for our hypothesis. The intended dm_meltingpot prisoners_dilemma_in_the_matrix substrate is a 2D-grid spatial coordination task that would have provided a non-trivial signal; it was unavailable in our runtime.